A New Statistical and Verbal-Semantic Approach to Pattern Extraction in Text Mining Applications

Dildre G. Vasques, Solange Rezende
University of Sao Paulo, ICMC
Sao Carlos-SP, Brazil, 13566-590
dildre.vasques@usp.br, solange@icmc.usp.br

and

Paulo S. Martins
University of Campinas, School of Technology
Limeira-SP, Brazil, 13484-332
paulo@ft.unicamp.br

Abstract

The discovery of knowledge in textual databases is an approach that basically seeks for implicit relationships between different concepts in different documents written in natural language, in order to identify new useful knowledge. To assist in this process, this approach can count on the help of Text Mining techniques. Despite all the progress made, researchers in this area must still deal with a large volume of information and with the challenge of identifying the causal relationships between concepts in a certain field. A statistical and verbal semantic approach that supports the understanding of the semantic logic between concepts may help the extraction of relevant information and knowledge. The objective of this work is to support the user with the identification of implicit relationships between concepts present in different texts, considering their causal relationships. We propose a hybrid approach for the discovery of implicit knowledge present in a text corpus, using analysis based on association rules together with metrics from complex networks to identify relevant associations, verbal semantics to determine the causal relationships, and causal concept maps for their visualization. Through a case study, a set of texts from alternative medicine was selected and the different extractions showed that the proposed approach facilitates the identification of implicit knowledge by the user.

Keywords: knowledge visualization, knowledge discovery, knowledge modeling, knowledge representation, implicit relationship, knowledge acquisition, causal concept maps

1 Introduction

The amount of knowledge accumulated in documents written in natural language represents a potential source for new discoveries. Due to the advancement of digital media, this knowledge can be stored and made available to those who need it [1]. However, researchers must deal with a significant number of publications, the overload of information, and the lack of data structure. This scenario promotes many challenges related to the acquisition, representation and analysis of this content [2].

To deal with these challenges, different models, processes, and methodologies have been proposed. Among these is the Knowledge Discovery in Text (KDT) [3, 4, 5], an approach used to find latent rules and patterns in texts that result in useful knowledge [4, 6, 7]. Due to the potential of the KDT process, it is fundamental to develop models that facilitate the process of discovering implicit relationships. Note that an implicit relationship, as the name implies, is a type of relationship between concepts that is hidden. In this work, we assume that it either exists as a proven or accepted concept or as a hypothesis to be verified. To make this process viable, faster and more efficient, one can count on the support from Text Mining (TM) techniques [8, 9, 10, 11]. The TM process is basically divided into five main phases: identification of the problem to be addressed, preprocessing of texts, extraction of patterns, post processing and use of knowledge [10]. TM fundamentally relates to the discovery of useful information from

unstructured or semi-structured documents. Its techniques enable automatic analysis of a large textual set with little manual intervention.

However, the area surrounding TM must still deal with the complexity inherent in the overall structure of manipulated texts since these documents are composed of complex and heterogeneous data. Therefore, it is essential to identify and consider existing relationships in the dataset [12] to better understand the general context of the processed texts. Despite the available technology, the extraction of knowledge from texts remains an arduous task [13]. Textual documents have a structure that requires the application of specialized techniques due to the implicit meaning assigned to each word in the human language [14] and between the relationships in which those words are involved.

The main advantage of a set of associated words is the fact that it relates to the context of the document [15], and it contributes to the maintenance of its representativeness. Association analysis is one of the several TM techniques presented in the literature [16], which is responsible for highlighting indirect relationships and making explicit potentially useful connections between the words that make up a set of documents. In an association analysis, the discovered relationships are represented by association rules. These rules make explicit the items that are most often listed in a database. Thus, in relation to TM, the association process aims to elicit existing indirect relationships between words, which may explain potentially useful connections. However, some of the discovered relationships are potentially false, because they may happen simply by chance. Therefore, the results of an association analysis must be carefully interpreted, since inference by an association rule does not necessarily imply causality. Causality requires the knowledge of cause-and-effect attributes in the data [12].

In this scenario, the detection of causal relationships could represent another method to determine potentially useful relationships, representing an interesting way of analyzing and understanding how the words (concepts) relate to each other. However, mining these relations is a challenge due to the difficulty of capturing the causality between events. In comparison with statistical coexistence, causal relations have some unique properties. For example, by saying that an event X causes an event Y, it is expected that Y will be influenced by X, not the opposite. This fact indicates that there is an intrinsic asymmetry between related concepts, which can not be disregarded. Therefore, there is a need to use approaches that explain how these cause-and-effect relationships occur [17].

In order to obtain such explanations, we can count on the support of techniques based on linguistic analysis. These techniques can use textual analysis of the syntactic and/or semantic type. Syntactic analysis is related to the superficial aspects of language, that is, its structure and form. It works as a set of combination rules to form words, phrases, and sentences. The semantic analysis, in turn, is related to the meaning of these constructions. We note, however, that several researchers are only committed to the syntactic function of a word in a sentence, or else, at the semantic level, they only focus on the meaning of words and their hierarchical (taxonomic) relationships [18, 19]. In most cases, the methods frequently used in the KDT process disregard the contextual semantic of the processed texts and, therefore, they provide few elements corresponding to the general content of the text.

The most common semantic approach used in TM is the treatment of latent semantics, performed using the Latent Semantic Indexing (LSI) method [20, 21], followed by applications related to the use of concepts or topics present in the documents [22, 23]. However, these approaches are not able to provide semantic explanations linked to relationships that do not belong to a semantic hierarchy taxonomic relationship. In order to obtain semantic explanations related to relationships between concepts of distinct semantic classes (non-taxonomic relationship), the analysis should be based on events. Thus, the main focus of this type of analysis necessarily falls on verbs. Verbs lead to the specification of words that fit in a sentence, the morphological forms of these words, the order in which they should appear in the sentence [24, 25, 26, 27] as well as their cause-effect relationships [28]. The contextual semantic based on events is related not only to the order (i.e. syntax) in which the words are distributed in a sentence but also to the semantic role (thematic role) that these words play in that sentence and how the set of all sentences intertwine to form the general context of a document. Thus, the aim of thematic role work is to identify the predicate (verb) of a sentence and then assign to its arguments (subjects, verbal complements) certain semantic roles such as agent, patient, instrument, place, among others. Therefore, semantic research techniques focused on the verb can help in the understanding of contextual semantics.

However, in addition to the problem of textual data complexity, the TM process must also deal with a large number of words present in a textual dataset. To deal with this issue, statistical measurements from complex networks may be used. There is a wide range of metrics in the complex networks approach available for the characterization of the topological properties of a linguistic network. These metrics can aid in the process of assigning weights to the main words and relationships present in the data. In this sense, metrics are considered as a type of weight assignment, in order to select a subset of more representative words. By using this approach, we can apply a filter to select valid and useful relationships for the discovery of knowledge. The interest in modeling and analyzing text as complex networks has been on the rise in recent years and considerable research in this area has already been accumulated [29, 30]. These investigations demonstrate that several aspects of natural language systems may be represented as complex networks. Their vertices depict linguistic units (such as words, morphemes, and phonemes), while links shape their relationships [31, 29].

Within this context, the goal of this work is to facilitate the inference of implicit relationships in a textual corpus,

by applying a TM technique based on association analysis, plus complex network metrics and verbal semantics to allow the visualization of explicit causal relationships in a concept map.

This approach allows us to add the advantages of the quantitative and qualitative analysis models to the TM processes. The extracted relationships may contribute to the discovery of implicit relationships by the user assisting in the KDT process. In essence, our methodology is based on association analysis that enables the identification of frequently associated words in the entire corpus (set of processed texts). These words and their relationships are then represented in the network format and are analyzed using quantitative measures from complex networks that provide a ranking of the concepts (nouns) that stand out in the set. With this ranking, we can decrease the volume of data and extract a subset of more representative concepts. With these main concepts and their explicit relationships, we model a causal concept map with the aid of linguistic techniques based on verbal semantics. This map enables the user to manually infer implicit relationships.

This work is structured as follows: Section 2 discusses background and related work. Section 3 shows the proposed approach. Section 4 addresses the implementation of the approach and the results. Finally, Section 5 presents the conclusion.

2 Background and Related Work

In this section, we review key concepts and related work on the three techniques used in this work, i.e. association analysis, complex networks, and verbal semantics.

2.1 Association Analysis

Bhardwaj and Khosla [16] point out that in TM, one of the main components to be considered in the treatment of textual data is context analysis. According to the authors, the techniques of association analysis stand out in obtaining this type of information. In addition, association analysis can be used to decrease the volume of data. This type of analysis is used to find patterns that describe highly associated characteristics within the data [12]. Thus, one can find useful relationships that, apparently, are not visible. Discovered relationships can be represented in the form of association rules. Thus, association rules extract the relationships between items (words) in a database, where A $\rightarrow$ B, meaning that when a term A occurs, a term B also tends to occur [32]. Thus, regarding TM, the association analysis aims to evidence existing indirect relationships between words, explaining the potentially useful connections. The strength of an association rule is measured in terms of support and confidence. Support determines the frequency with which a rule is applicable to a particular dataset. Confidence, in turn, determines how often items in a set appear in transactions that contain another set of items [12].

Support and confidence measures are the most common in extracting association rules. However, this model generates a large number of rules, which may render the process of user analysis difficult [33]. Moreover, in many cases, the rules obtained using only these two measures may present many false relationships, even with high confidence values [34]. This is especially true when the antecedent support value is low and the consequent support value is high, so that most of the extracted rules are composed of items that most occur in transactions.

Considering the examples illustrated in Table 1, the values shown in the second column and the third column represent, respectively, the support value of each word. For example, the word vitamin, independently of the rule it may belong to, has a support value of 29.3 (see the first through the fourth lines where it occurs). The values presented in the fourth and fifth columns represent the support and confidence of each association rule, respectively. We conclude, by analyzing these results, for example, that the "risk $\rightarrow$ vitamin" rule may represent a false relationship, since the support value of the antecedent (risk = 2.8) is low compared to the support value of the consequent (vitamin = 29.3). Therefore, in this case the rule that would most accurately represent the relationship between these two concepts is the "vitamin $\rightarrow$ risk" rule, where the antecedent has a higher support value (vitamin = 29.3) and the consequent a lower one (risk = 2.8).

The selection of the rule in this case may be confirmed by its confidence value. In this case, "risk $\rightarrow$ vitamin" has a low value (9.7), whereas "vitamin $\rightarrow$ risk" has a high value (28.6), which leads to its adoption. Thus, for a rule to be considered strong, it must have adequate support and confidence values. The decision about which rules to keep and which to discard during the mining process is based on the values of these two measures. This means that support and confidence act as key measures of interest in the association rule mining process. However, this process is not trivial, and in many cases rules that appear to be classified as false may reveal relevant information to the user. This problem was worked out in our approach by using metrics from complex network theory.

Ruiz et al. [35] used association rules to merge information extracted from various datasets in order to decrease the volume of data, their distribution and volatility. The model proposed by the authors produces meta-association rules, that is, rules in which the antecedent or the consequent word may also contain rules, to find joint correlations between the trends found individually in each dataset. d'Amato et al. [36] have shown that the use of association rules can also help in the coupling between ontologies and statements that may be out of sync. The authors proposed a method

Table 1: Association Rules extracted from the database.

Rule	Support Antecedent (LHS)	Support Consequent (RHS)	Support Rule	Confidence Rule
"vitamin → pregnant"	29.3	9.9	4.3	42.9
"pregnant → vitamin "	9.9	29.3	4.3	14.6
"vitamin → risk"	29.3	2.8	2.8	28.6
"risk → vitamin"	2.8	29.3	2.8	9.7
"labor → premature"	5.7	5.7	4.5	80.0
"premature → labor"	5.7	5.7	4.5	80.0
"risco → labor"	2.8	5.7	3.4	60.0
"labor → risk"	5.7	2.8	3.4	34.3

to discover a multi-relational relationship through the use of association rules, considering intentional knowledge. Moreover, the rules discovered could be directly integrated into the ontology, enriching its expressive power and increasing the assertive knowledge that could be derived.

Amin et al. [37] used association rules to assist in the development of a dengue control system in Pakistan. Through the rules obtained, the authors have shown that if a virus attacks somewhere, it is possible to predict its next geographic target. In Vasques et al. [38], association rules were used to help telecom companies with the adoption of business strategies to satisfy and maintain their customers. Considering the reports of occurrences of telephone service, which were registered in a private database, the authors, through the rules generated, identified the profiles of the users who remained faithful to the operator and the profiles of those who left them. The resulting association rules allowed the understanding of the reasons that led them to stay or give up the operator.

2.2 Complex Networks

In addition to the body of work that uses association analysis, we also find in the literature work from complex networks for the exploration of the contextual aspects of a document, currently, as an alternative to representations based on the vector space model. Networks are composed of objects and their relationships. A network can be described as a graph and, as such, inherits its conceptual properties. A graph is defined as a mathematical structure consisting of two sets: V (vertices) and E (edges), resulting in the formula G = (V; E). A certain vertex is adjacent to another if there is an edge that joins both. A graph can also be directed or not. In a directed graph, also called a digraph, there is no symmetry between the two objects of the relation. The graphic representation of the edge is an arrow, which points out the relation of one concept to another. In an undirected graph, there exists a symmetrical relationship between the edges.

Networks are able to represent different types of objects and different types of relationships [39] and obtain better results in comparison to algorithms based on the vector space model [40]. Such models are used to represent relations between textual documents or between the many concepts that make up the documents. The analysis based on complex networks allows the extraction of different metrics, such as the average degree of vertex, hubs, betweenness centrality, and clustering coefficient, among others. [41]. Wachs-Lopes and Rodrigues [31] developed a study that proved that complex networks also provide a useful tool to aid in the understanding of context-related issues. In their study, the authors modeled two complex networks, the first being in English and the second in Portuguese, and they presented the study of the dynamics of these two networks. They showed the behavior of the small world and the influence of the hubs, thus suggesting that these databases have a high degree of modularity, which indicated specific contexts of words. Ke et al. [42] used complex networks to develop a method to analyze the quality of nonlinear Chinese-language text. For the development of the study, the authors used texts produced by university students in China. These texts were then represented as free-scale networks (word adjacency model), from which typical network resources were obtained, such as clustering coefficient and network dynamics. The results revealed that complex network features of different text qualities can be clearly revealed and used in potential applications for other text analysis.

A network representation can also increase the number of links that make up the intermediate path between different words, allowing the extraction of all the words involved in a relationship. The relations between words enable the extraction of patterns that would hardly be captured by algorithms based on the vector space model, besides being useful to improve the quality of the extracted standards [43]. According to Solé et al. [44], different configuration models can be used to generate graphs that represent the content of documents written in natural language. These templates can be configured to generate co-occurrence networks (words that appear together in a sentence), syntactic networks (words that have syntactic dependencies) and semantic networks (words linked by semantic relations).

Xu et al. [45] developed *Knowle*, an online news management system, in which they introduced a semantic link network model. The central elements of *Knowle* are Web news events linked by their semantic relationships. Knowle

is a hierarchical data system, composed of three different layers formed by concepts, resources, and events. In short, this system provides the various semantic relationships between these layers. In their case study, the system was used to organize and mine health news, and it showed its potential to analyze large databases in the health field.

Thus, in order to obtain a richer representation of texts written in natural language that can be processed by computer systems, some methods for the generation of different types of networks from texts were created. In this model of representation, usually, vertices represent the nouns (subjects and objects) of the sentences and the edges represent the verbs, thus forming a graph. As an example of this model of representation, we find in the literature the semantic networks of Sowa [46] and also in the Resource Description Framework (RDF), whose model is the basis of the Semantic Web. In RDF, the representation is done in triple format (subject-verb-object), and the set of several triples on the same subject is seen as a directed graph [47].

Recently, also in the area of health, SemMedDB, the Semantic MEDLINE [48], was converted into a graph database called Neo4j, which facilitated the extraction of relations using patterns of discovery and enabling the fusion information sources. SemMedDB is a repository of semantic (subject-predicate-object) extracted by SemRep [49]. SemMedDB currently contains information on approximately 94 million predictions of all PubMed citations (about 27.9 million citations, on 31 December 2017) and forms the backbone of the MEDLINE Semantic application [50, 51].

This body of work reveals the importance that association analysis and complex networks approach assume in the context of KDT. However, for the most part, the methods used still have little emphasis on contextual semantics, thus impairing the performance in terms of the representativeness of explicit relationships and the discovery of implicit relationships. Thus, we note that a large part of the recent KDT and TM research has focused on reinforcing more advanced semantic applications.

2.3 Verbal Semantics

The semantic aspects that one wishes to emphasize in a given representation guide the type of extraction and selection of certain types of semantic relationships between words [52, 53]. There are two types of semantic relationships existing between the words (concepts) that guide the rules of interpretation of their meanings: taxonomic and non-taxonomic relationships. A taxonomic relationship is one that occurs at the word level, and it includes items of the same category (class of words), such as the relationship between the words “chair - table - armchair”, which fall into the class of “Furniture”. Therefore, this type of relationship refers only to the lexical structure of terms and concerns the practices of naming, defining and categorizing them [54, 55, 56]. The non-taxonomic relationship is that which occurs at the sentence level and is also called thematic relations (or semantic relations). This type of relationship includes words that are related to scenarios or events [26]. In this type of relationship, words assume certain “thematic roles (or semantic roles),” that is, their functions in the sentence, such as “agent”, “patient”, etc. In this way, the words present the non-taxonomic relationships share an associative relationship of distinct classes [26, 57, 58, 19]. The relationship between the words “Joiner - builds - table” reveals the association between a class referring to “Professions” and a class relating to “Furniture”, linked by means of a verb.

The discovery of non-taxonomic relationships was identified in the literature as being the most difficult to obtain [59, 2]. This difficulty is due to the linguistic complexity involved in the construction of a sentence and the identification of the function (thematic role) that each word (concept) plays in a sentence. The assignment of thematic roles to non-taxonomic relationships is difficult, since various relations between instances of the same general concepts are possible at semantic-lexical and sentence level [60, 2]. Besides that, there is a lack of consensus in the literature on the actual number of thematic roles, as well as the lack of consensus on the terminology used [28]. Several approaches have been proposed, which differ from one another, generating a large number of “semantic types” for the concepts [61, 62, 63, 64, 65, 66].

Nevertheless, this body of work regards the verb as the central element for the discovery of non-taxonomic relationships. Thus, the verbs gained prominence in semantic investigations [60, 59, 67] and began to be used in several types of research also as an instrument for the extraction of causal relationships in textual documents. Garcia et al. [68] developed COATIS, an automatic system created to acquire knowledge of causality from texts. Khoo et al. [69] developed a type of automatic extraction of information from and made from texts. For this, they used linguistic clues without the use of any domain knowledge. These clues were based on causative verbs, constructions (“if ... then ...”) and in adverbs and causative adjectives. Girju [70], presented an intensive system of supervised knowledge that is based on constructs structured into semantic (or triple) predications of “noun-verb-noun”. These triples were used to identify new pairs of nouns that codify causes and effects (e.g, “Tsunamis cause tidal waves”). In that system, the verb is identified from a set of 60 causal verbs present in Word-Net (for example, “cause”, “leads to”, “provokes”). Thus, in this work, the author developed an approach to automatically identify lexical-syntactic patterns that express the causal relationship and semi-automatically validate the patterns.

A disadvantage in the research papers cited above is the lack of a context of sentences, which are analyzed in isolation. This generates a loss of information, which may be fundamental to the understanding of knowledge. Most systems are methods to create non-taxonomic relationships (based on in events) but do not consider the context in which those relationships may be inserted. The context is related to the comprehension of the set and interlacing of the

various sentences that make up a text, not just the comprehension of some isolated phrases. With one chain of logical links between words and sentences, explicit relationships can be understood and hidden relationships can be identified. According to Nastase et al. [71], causality is a type of complex relationship and not primitive, which can be refined into more specialized sub-relationships. In this sense, only the labeling of the concepts (thematic roles) is not enough to reveal all the relationships existing causal relationships in a text. It is also necessary to cross-reference the labeled sentences to obtain the more specialized sub-relationships that will provide sufficient information for the identification of causal relationships. Hashimy and Kulathuramaiyer [72] sought to consider the context of the input text to obtain a semantic classification of relationships, in particular, relationships of causality. They proposed a framework to extract the initial semantic causality relationships of the input samples. They then filtered these patterns using two algorithms.

These papers contributed significantly to demonstrate the importance of more complex event-based semantic analyzes to logical understanding of the relationships existing between the different concepts. They also demonstrated that the identification of causal relationships can contribute to more advanced TM. However, most of the time, these papers used lists of predefined verbs, neglecting any other relationship that did not derive from that list. This fact generates a loss of relevant information, especially with respect to the context and the discovery of implicit relationships. Another problem is the limitation related to triple extraction model, which considers only the subject and the complement object (usually direct object) of the verb, neglecting the other linguistic information present in the verbal syntagma (as adverbial adjuncts, for example), which are fundamental for obtaining an understanding of meaning.

Meaningful words or sentences embedded in any paragraph of a full text can be processed to extract logical relationships. With a chain of logical connections, hidden relationships can be identified. From the identification of semantic relationships between words, it is possible to identify the logical meaning of such relationships, without the need for explicit words such as “cause” or “effect” [73]. This idea was also explored by Vasques et al. [28, 74] who, with the help of verbal semantic analysis, developed a process called Verbka that is capable of extracting causal relationships from a text. To this end, a basic distinction between a participant X (Cause) and one (or more) participant(s) Y (Effect) is necessary.

To assist in the process of obtaining more meaningful causal relationships, Vasques et al. [74] developed a process of knowledge acquisition called Verbka. This process considers, together, the approach of semantic roles, verbal semantics and the context of a document for the identification of causal relationships. Based on theories of Cognitive Semantics that approach takes the verbal semantic as the central unit of discourse [61, 75, 76] and provides a relevant linguistic theoretical basis for sustaining a process of knowledge extraction based on cause-and-effect. The classification of concepts in thematic proto-roles proposed by Dowty [76] makes it possible to reduce the large number of semantic roles present in literature in only two macro roles (proto-roles), classified as X and Y (proto-agent and proto-patient). The Langacker [75] model of energy flows is harmonized with the theory of thematic proto-roles since the focus of this theory is also the definition of an agent (responsible for causing an action) and of a patient (affected by an action).

Therefore, the process requires that a text must be initially fragmented in linguistic pieces. Then, this fragmented text will be finally reconstructed in the form of concepts and their causal relationships, which will then be interlaced and represented in the form of a graph (network), composing a causal concept map. It should be noted that in this process all the verbs present in the texts are considered. Each of these verbs is then inserted into a specific verbal category (verbal semantic), according to its typology: cause-and-effect verbs, reflexive verbs, and static verbs. In this approach, the conceptual modeling of propositions extracted from the text is suggested in a concept map format because of its ability to represent knowledge in an interactive and dynamic model, facilitating the user’s visualization and comprehension [28].

2.4 Comprehensive Work

In this subsection, we review work that combines one or more of the previous techniques. Salloum et al. present a study that demonstrates a comprehensive view of text mining and its current research status. The authors point out that there is a limitation in the treatment of information extraction from research articles using data mining techniques. According to Jabri et al., The textual information user must deal with the exponential increase of data and the difficulty of structuring and understanding them. This creates a challenge for information retrieval. In order to deal with these problems, different authors have been using the task based on extracting association rules.

Liu et al. [77] presented a mining approach to detecting topic relationships based on parallel association rules, called PARMTRD (Parallel Association Rules Multi-Topic Relationship Detection). This approach detects relevance across multiple topics by selecting and assembling association keywords in keyword sets, which help you find event sources across multiple topics. Li et al. analyzed high-frequency word co-occurrence networks in the bioinformatics literature, in addition to their structural and evolutionary characteristics. They analyzed the co-occurrence relationship between the high-frequency words of the articles and created a co-occurrence network of these words. The results showed that the high-frequency word co-occurrence network in the bioinformatics literature was a small-world network with non-scaled distribution.

Zupanc and Davis [78] investigated large automatically constructed knowledge bases (KBs) in order to learn how these knowledge bases generate new knowledge in the form of inferred facts or rules that define regularities. They proposed a scoring function to evaluate the quality of a first-order rule learned in a KB. The proposed metric was intended to include information about tuples that were not contained in KB when assessing the quality of a potential rule. However, one of the major challenges of TM is related to the semantic aspects of the texts. In this sense, researchers have presented works focused on improving the semantic representation of information and knowledge extracted from texts.

Kralj, Robnik-Sikonja, and Lavrac [79] proposed a semantic data mining approach called NetSDM, which transforms available semantic knowledge into a network format, followed by classification. and removing nodes based on network analysis to reduce the size of the original knowledge. In the experimental evaluation of the proposed methodology on acute lymphoblastic leukemia and breast cancer data, they demonstrated that NetSDM achieved improvements in time efficiency and that the rules learned were comparable or better than the rules obtained by other semantic data mining algorithms. Biryukov et al. implemented a pipeline that, while exploring text mining and semantic technologies, to help researchers access the semantic content of PubMed and Elsevier's full-text abstracts and articles. Relationships are extracted along with the context that specifies the location of detected events, preconditions, temporal and logical order, mutual dependency, and / or exclusion. To assist in the task of implementing improvements to semantic aspects of textual representation, several authors have used the network science approach.

Sizemore et al. [80] created semantic resource networks, where words correspond to nodes and connections correspond to shared resources to understand language learning and knowledge acquisition. According to the authors, nodes and their connections characterize the structure of strongly interrelated word groups. Using the network topology, they found that, despite varying word order, the overall organization remains similar. They also showed that network measurements correlate better with sparse parts than basic lexical properties. Raimbault [81] introduced a methodology combining citation network analysis and semantic analysis. They collected a corpus of about 200,000 articles with their abstracts and the corresponding citation network that provides a first citation classification. They extracted relevant keywords for each article through text mining, building a semantic classification. They studied the qualitative patterns of relationships between endogenous disciplines within each classification and showed the complementarity of the classifications and their associated interdisciplinary measures.

Yang et al. [82] used a Subject-Action-Object (SAO) network to analyze trends in the development of graphene technology. The authors, by means of a network representation, represent the subject and the object of the sentences as nodes and the action as the edge. An empirical study of graphene technology was used to apply the method, employing concepts of structural holes, changes in node degree distribution, and changes in network centrality to detect trends in the development of graphene technology. Vlietstra et al. [83] have shown that semantically integrated knowledge, drawn from biomedical literature and structured databases, can be used to automatically identify potential migraine biomarkers. To do so, they filtered out composite biochemical concepts and classified them by their potential as biomarkers, based on their connections to a subgraph of migraine-related concepts. The minimum unit of knowledge in the network is a triple formed by two concepts linked by a predicate (subject - predicate - object). The sources (origins) of each triple are also included in the network.

Qiu et al. [84] feature a multiphase correlation search framework to automatically extract non-taxonomic relationships from big data documents in a smart city. Different types of semantic information were used to improve system performance. They proposed a method based on semantic graphs to combine semantic graph structure information and term context information to identify non-taxonomic relationships. They then used different semantic types of verb sets based on syntactic dependency information, classified to act as non-taxonomic relationship labels. Yu et al. [85] developed an approach for extracting relationship information from bioentities present in texts, in order to label certain relationships as true or false. To this end, they used Natural Language Processing (NLP) and a theoretical-graphic algorithm. The method, called GRGT (Grammatical Relationship Graph for Triplets), extracts the pairs of terms that have certain relationships and also the type of relationship (the word that describes relationships). Thus, they represent the sentence structure as a dependency graph where words are nodes and edges are typed dependencies. The shortest paths between the word pairs in the triplet are then extracted, forming the basis of the information extraction method.

Our work aims at the syntactic and semantic exploration of sentences for more detailed information and a more meaningful analysis of the extracted relationships [21, 86, 87, 88]. The process of extracting implicit relationships used in our work, in addition to having techniques for extracting association rules and metrics from complex networks, uses as attributes the semantics contained in the verbs that make up the existing relationships between concepts. This type of semantic analysis allows the understanding of causal relationships existing between the concepts of the texts analyzed. Thus, it seeks to maintain, as far as possible, the contextual semantic load, in order to use the relationships extracted in knowledge discovery. It may be seen from the literature that there is early work that uses association rules, complex networks, and work that combines both. To our knowledge, no prior work has combined association rules, complex networks and concept maps, structured around verbal semantics, to support the task of uncovering implicit relationships in textual databases.

3 Proposed Model

Much of the information produced and made available by digital means is in textual format. The texts do not present data organized in rows and columns and, therefore, they are composed of a set of unstructured data. Because it does not have a formal structure, this type of information becomes difficult to analyze. To assist in the process of extracting relationships from this type of information, we can rely on a set of techniques from different areas, which may contribute to a more advanced TM. However, the KDT process is not trivial, mainly due to the lack of methodologies that emphasize the contextual aspect in which the extracted relationships are involved. There are two different types of approaches that can be used to analyze textual data: statistical analysis (based on the frequency of terms) and semantic analysis (based on the functionality of terms). These approaches can be used either separately or together [89]. In statistical analysis, the importance of a term is calculated by counting the number of times it appears in the text. Its process is centered on statistical learning from data usually involving the steps of data coding, data estimation, and document representation models. The semantic analysis (linguistics) is based on Natural Language Processing (NLP) techniques [90, 91], which is used to evaluate the sequence of terms in the context of the texts in order to identify their function, i.e. it's meaning in a given context.

Thus, the present work proposes a process for discovering implicit relationships present in a text corpus, using a hybrid approach based on three different techniques: association analysis, complex networks metrics (which are statistical), and analysis based on verbal semantics (which is linguistic). Fig. 1 illustrates the phases (and their respective steps) that make up the proposed TM process for uncovering implicit relationships. These phases enable the interconnection of textual content represented by concepts (words) in a domain of analysis with the purpose of supporting the process of knowledge discovery. The first phase is related to problem identification. It must be understood what is the purpose of the TM, for example, "*Why some emerging infectious diseases cause epidemics*". The second phase concerns the texts preprocessing, including usual tasks that prepare the text for the extraction of terms. The third phase consists of deriving patterns (i.e. causal relationships) that summarize the underlying relationships in the data. It extracts and explains the results obtained through the visualization (through a concept map in our case), considering that a different view of textual data may reveal important features. Pos-processing is the fourth phase, and it is necessary to evaluate and interpret the patterns extracted from the documents. Finally, in the last phase, after being evaluated and validated, the extracted knowledge may then be used by the users in a way that suits them (Knowledge Utilization).

To test the feasibility of the proposal, we apply the process to a set of ten health texts dealing with alternative medicine, which are processed and represented in a network format in order to enable the exploration of data from different points of view. We highlight that one of the major challenges of TM is the high dimensionality of the words. A set of documents may have hundreds or even thousands of words. However, many of these words are irrelevant and do not contain useful information for the mining task. An option to deal with this issue is to analyze the words in terms of their relationships. In this type of approach, the first step in the analysis is to find relationships between the different words to then execute the remaining analysis using these relationships, rather than analyze the words themselves [12]. Thus, the representation model used to visualize the data is a causal concept map, since it is able to capture and reveal important aspects of the relationship between the words (i.e. concepts) under analysis. The next subsections address each phase and the corresponding steps.

3.1 Phase 1: Problem Identification

This step is fundamental to structuring the rest of the mining process. The focus of this phase is to detect a problem and select the database and techniques that may possibly help in its resolution. At the end of the entire mining process, it will, therefore, be necessary to check whether the initial problem has been solved.

1. *Define Goals*. It is at this stage that the objective of the mining must be delimited and specified, having clear what is expected from the analysis of the data. It is also at this stage that one selects the base of texts that will be explored and define how the results will be used.

2. *Select Textual Documents*. This step consists of selecting the textual documents that will compose the input dataset.

3.2 Phase 2: Preprocessing

Preprocessing transforms documents written in natural language into a viable data format, which is then used to extract interesting relationships and information relevant to the KDT process. To improve the quality of mining standards and the time required for mining, some data cleaning techniques must be applied in the preprocessing phase.

1. *Prepare Text*. it is necessary to prepare the documents in the collection to be processed by the algorithms used for pattern extraction. Therefore, it is first necessary to generate the attributes that represent the textual

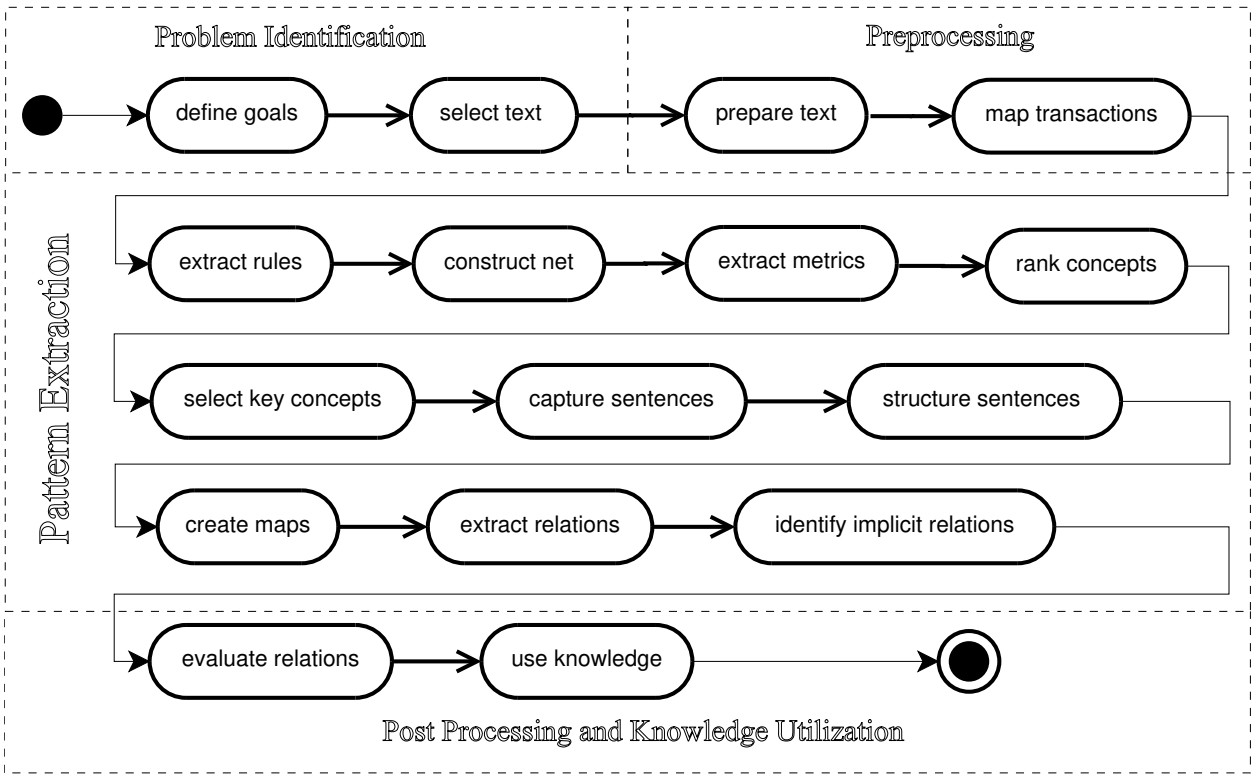

Figure 1: Approach for uncovering implicit relationships.

documents. To this end, treatment, cleaning and volume reduction activities are usually performed. The cleanup aims to remove special characters (digits, punctuation and accent marks, and line breaks), plus some stop-words (articles, prepositions, conjunctions). After obtaining the attributes that represent the textual documents, a representation of the data is then generated in a format that is suitable for the extraction of knowledge.

2. *Map Documents into Transactions*. After the preprocessing phase, the set of texts is mapped into transactions (different sets of items). In association analysis, a collection of items is called an item set. In the case of TM, these items are the words that make up the texts. The mapping of the textual document into transactions allows the extraction of association rules. Transactions are sets of words (items) under the same context of analysis. They may be obtained either from sentences, paragraphs or from a sliding window. The paragraph mapping considers the set of all the words that are contained between the punctuation marks ".", "!", or "?", followed by a line break. The mapping in sentences considers the words that appear between the ".", "!", or "?" signs. To obtain the transactions in this work, we opted for the sliding window technique. [92]. In a sliding window mapping, it is possible to limit the maximum distance that two words should appear in the text to be considered as their co-occurrence. Thus, the first transaction contains only the first word of the document, the second contains the first two words, and so on, until the window contains the number of words equal to the size defined for the window [92].

3.3 Phase 3: Pattern Extraction

This step aims to extract knowledge through the application of pattern extraction techniques. These techniques are based on the main tasks of Data Mining (classification, clustering, regression, and association). The tasks to be performed are defined according to the ultimate goal of the knowledge extraction process. It is possible to summarize the main activities related to the extraction of patterns in predictive and described activities. Predictive activities (classification and regression), which consist of the use of supervised machine learning algorithms, which require a set of training examples that have meta attributes. Descriptive activities (obtaining association rules, clustering, or summarization of document collections) consist of the use of unsupervised machine learning algorithms that detect intrinsic behaviors in the collection of data that do not have meta attributes. In this work, we used the descriptive activity, using association rules, plus metrics derived from the analysis based on complex networks and semantic information obtained through analysis based on verbal semantics. In summary, our goal it to identify relationships and thus we use association rules for textual documents, which encode important information about the relationships

between items (i.e. words, documents, etc..) and, ultimately, are able to identify patterns, themes, and context [93]. Thus, in this work, we use the association analysis with the objective of (i) identify the words that are most frequently related; (ii) decrease the volume of data; (iii) provide pairs of related words for the construction of a complex network.

The approach based on complex networks can provide important metrics to assist in the selection of words. In this work, we use complex networks to (i) extract metrics from the network; (ii) use metrics to weigh words; (iii) rank the core words of the network; (iv) extract a subset of words; (v) decrease the number of words.

The verbal semantics may help in the detection and understanding of causal relationships between concepts. Thus, we use a knowledge acquisition process (Verbka) based on verbal semantics to: (i) select the main concepts (main nouns) ranked in the previous step; (ii) search the original textual dataset for the sentences in which those concepts appear; (iii) divide the sentences into linguistic blocks using Verbka to extract logical relationships; (iv) create a causal concept map to enable the visualization of the mined result; (v) add verbal semantics to the explicit relationships; (vi) make inferences and find implicit relationships.

In order to achieve these objectives, the following steps should be taken:

1. *Extract Association Rules*. The execution of this step is carried out with the support of the *Features gEnerator* tool [92]. An association rule is represented as an implication in the LHS $\rightarrow$ RHS format, in which LHS and RHS are, respectively, the Left Hand Side (antecedent) and the Right Hand Side (consequent) of the rule [94]. In order to evaluate the strength of an association rule, two classical measures are used: support and confidence. The value of the support is applied to measure the co-occurrence strength between LHS and RHS and does not indicate possible dependencies between RHS and LHS. The confidence measures the strength of the implication described by the rule [95]. The problem of obtaining association rules can be decomposed into two steps: 1) find all k-itemsets that have support greater than or equal to the minimum support specified by the user (minSup), and 2) use the frequent k-itemsets, with k $\geq$ 2 (greater or equal), to generate the association rules [32].

2. *Select Association Rules*. This step is carried out by the user. For the development of this work, we select all the rules composed of two words, that is, one word that forms the antecedent and another word for the consequent.

3. *Construct Complex Network*. With the association rules selected in the previous step, we may build the complex network. In a simplified way, the concept of a complex network is that of a set of elements that are connected to each other [96]. Networks (graphs) are able to model textual content and may be useful in TM steps [39]. In this phase of development of the work, we use a non-directed graph because the direction of the relationships will be considered only in the Semantic Analysis steps.

4. *Extract Metrics*. Metrics extracted from networks can also be considered as an automatic weight assignment type for the objects that make up the dataset [12]. This weighing is a key element is this work, as it is the alternative used to select the words that must be either maintained or eliminated from the set of texts. Heavily connected words must be selected because they play an important role in the overall context of documents. Thus, the selected words and their relationships will compose a subset of data. This subset enables us to reap the full benefits of reducing the number of words. Complex networks have a set of definitions that can describe a series of behaviors related to a network, such as the degree of a vertex, its hubs, the average connectivity, the shortest path, its diameter, clustering coefficient, closeness centrality, betweenness centrality, among others. To this end, we use the Gephi Tool [97], a free software available at (https://gephi.org). It enables us to view and analyze networks by providing their various metrics. For the development of this work, we used the following metrics: 1) the *average degree* of a vertex (number of connections); 2) hubs (vertices with greater intensity of connections); 3) *betweenness centrality* (quantifies the participation of a vertex *i* in paths of minimum length), and 4) *clustering coefficient* (number of connections between the neighbors closest to a vertex) [41]. These metrics are used because they emphasize the most connected nodes. These nodes are closer to the underlying context of interest. Due to their larger connectivity, they may correspondingly uncover a larger number of implicit relationships than the nodes that are less connected. On the other hand, a remote node would reveal far less implicit relationships and these could be more outside the context of interest.

5. *Rank Concepts*. The metrics obtained in the previous step may help with the identification of the main concepts (nouns) of the network. The *Gephi* tool generates a ranking of the words (vertices) for each of the measures of interest, providing a list in descending order of corresponding values, which effectively contributes to the identification of the main concepts of interest, that is, the most representative concepts of the network built with association rules.

6. *Select Key Concepts*. Based on the word ranking obtained in the previous phase, it is possible to make a selection of two (or more) key concepts according to the users interests and application goals. The selected concepts serve as the basis for the construction of a subset of data.

7. *Capture Sentences*. With the selection of key concepts performed in the previous step, we capture (or retrieve) the sentences present in the original dataset that contain at least one of the previously selected key concepts.

8. *Structure Sentences*. Verbka decomposes a sentence into a set of minimal linguistic blocks formed by concepts. This is possible by applying a sequence of steps that input the preselected sentences and output a structured table of concepts and their relationships. These tables can provide a structured textual knowledge base, which can be constantly expanded with the insertion of new information.

9. *Represent Relationships through Concept Maps*. The table obtained in the previous step is then mapped into a concept map, which enables the visualization of the explicit relationships from the data subset. Concept maps graphically illustrate the relationships between the concepts of a knowledge domain [98]. Each concept is positioned within a circle or box (vertices) and relates to other concepts by means of arcs (edges) which represent the connections (verbs, prepositions). Concepts connected by an arc give rise to a proposition, which is the particular characteristic of concept maps [99].

10. *Extract Causal Relationships*. After the construction of the map, it is possible to assign semantics to the verbal relationships existing between the concepts and to create a causal concept map. The verbal relationships are divided into three categories which are represented by colored arrows: red arrows represent action verbs, i.e. causal relationships where the action starts from a "cause" concept and affects an "effect" concept , which in turn is modified by this action (e.g. she breaks the vase); blue arrows indicate reflexive relationships. In this case, an action departs from a cause concept and returns to the same concept (e.g. she washes herself, she feels happy); black arrows represent non-action verbs, indicating the absence of action and, therefore, absence of a causal relationship (e.g. she stays at home).

11. *Identify Implicit Relationships (User)*. Based on the visualization of the map, we can then proceed to the qualitative analysis. This analysis should be performed by humans as it requires an overview of the context. Therefore, based on this map, we look for connections that are not explicit in the network but can be inferred from the knowledge that already exists there. It is the cause-effect semantics present in the map and the visual aid (layout of information) that it provides that facilitates the identification of implicit relationships by humans.

3.4 Phase 4: Post-processing (Evaluate Uncovered Relationships)

In this step, it is necessary to evaluate and interpret the patterns extracted from the documents. Such as the other steps, this step should also be guided by the objectives defined at the beginning of the process. Certain aspects of extracted knowledge, such as representativity, novelty, validity, and applicability, should be evaluated.

3.5 Phase 5: Knowledge Utilization

After being evaluated and validated, the extracted knowledge can then be used by the users in a way that suits them.

4 Implementation and Results

In this section, we address the experiments that were carried out using the methodology described in Section 3. The problem of interest was to find out *how vitamins may affect the human gestation*. To this end, we decided to use a set of ten textual documents ranging from alternative medicine ("human gestation" and "vitamin" domains), in order to extract the implicit relationships in these texts. The texts were taken from the Cochrane website [1], which is a global and independent network of researchers, practitioners, patients, caregivers, and people interested in health, who work to produce reliable health information. The texts had an average of 550 words and the combination of all texts had 5575 words.

After selecting the texts, we proceed to the data preprocessing phase and the extraction of association rules with the support of the FEATuRE tool [92]. The texts were preprocessed separately, since the frequency of a given word within the whole collection may be negligible, whereas if processed separately, may not be. We opted for the removal of stop words to decrease the number of words. The texts were not subject to stemming in order to keep the semantic quality (related to the meaning) of the extracted words. In this work, we chose word frequency instead of stemming to generate transactions.

The documents were then mapped into a set of transactions with sliding windows of size five and step one (this size was chosen since it represents the usual number of connected words in a phrase). Sliding windows were chosen instead of sentences since they generate more transactions. The sliding windows correspond to excerpts from textual documents and they extract sets of words that are related to a specific context of the document. Table 2 provides an example of the transactions extracted.

For the extraction of association rules from the transactions previously obtained, we configure the tool for operation in *automatic support* mode. In this mode, the support value is generated by the tool. It takes into account the average

[1] (http://brazil.cochrane.org/)

Table 2: Examples of transactions extracted from a text.

...
{*supplements, vitamin, woman, health, produced*}
{*vitamin, woman, health, produced, body*}
{*woman, health, produced, body, human*}
{*health, produced, body, human, exposition*}
{*produced, body, human, exposition, excess*}
...

frequency of words in transactions, thus exempting the user from knowing the characteristics of the collection of documents. This avoids the definition of a low *minimum support* value, which could generate a large number of rules [92]. Table 3 shows some examples of extracted rules, where the first value represents the support and the second the confidence values.

Table 3: Example of association rules composed of two words extracted from transactions.

vitamin ← *supplements (3.4, 80.0)*
supplements ← *vitamin (3.4, 11.7)*
clear ← *is (3.4, 80.0)*
is ← *clear (3.4, 60.0)*
labor ← *premature (4.5, 80.0)*
premature ← *labor (4.5, 80.0)*
risk ← *premature (2.8, 50.0)*
premature ← *risk (2.8, 28.6)*
risk ← *labor (3.4, 60.0)*
labor ← *risk (3.4, 34.3)*
risk ← *reduce (4.5, 80.0)*
reduce ← *risk (4.5, 45.7)*

We considered the association rules only for sets of two words in order to preserve a common structure for the construction of the complex network in the next phase. In possession of these rules, we start to build the complex network using the Gephi tool. For network modeling, we consider each set of rules as related items. Thus, the network was created by mapping items as vertices of the network and the relationship between them was represented as non-directed edges. We emphasize that in this phase, the order of the relationship (edge sense) between the items does not interfere with the measures of interest. Upon execution, the tool generated a network as a non-directed graph of 161 vertices.

Networks have several properties that are part of their topology. Having built the network, we move on to the calculation of the metrics. After performing these calculations, it was possible to generate a ranking of items based on the metrics, i.e. degree of a node, clustering coefficient, hubs and betweenness centrality. Notice that the verbs must be excluded from the ranking, since 1) they represent the connections between the items that compose the causal concept map in the next phase, and 2) in this phase we wish only to capture the main nodes (items, i.e. nouns) that should compose the causal concept map. In a later stage, these verbs need to be retrieved back to complete the map. Table 4 shows the ranking of the average degree and hubs, and Table 5 shows the ranking of the betweenness centrality and clustering coefficient. We emphasize that the first items of the ranking are the most significant and are more likely to have implicit relationships due to their importance to the network.

An analysis of these tables shows that some of the most repeated concepts in all rankings are “vitamin”, “labor”, “risk”, “premature” and “pregnancy”. The most common approach to dealing with the number of words is to select a small subset of these words (usually two) for the display. This technique constitutes a type of dimensionality reduction [12]. For the sake of illustration, we select two concepts that henceforth are called key concepts: “vitamin” and “premature”. Note that this step admits some flexibility in that the user (analyst) may also pick up other concepts that he or she may be interested in, other than the ones selected here.

The next step of the process focus on obtaining (retrieving) all the sentences of the input corpus that have either one or both of the key concepts. Each retrieved sentence is then broken into linguistic blocks, as proposed by the Verbka process (Table 6). In Verbka, a number of steps are applied corresponding to some linguistic adjustments in the sentences, such as removal of articles, separation of coordinated sentences, and conversion of sentences from the

Table 4: Ranking of vertices using a) average degrees and b) hubs.

Vertices	Avg. Degree	Vertices	Hubs
vitamin	12	premature	8,87E+0.3
labor	6	risk	8,87E+0.3
pregnancy	5	evaluated	2,31E+0.3
supplement	4	study	8,14E+0.2
studies	4	hypothyroidism	8,14E+0.2
women	4	vitamin	0,575
orofacial	4	supplements	0,284
premature	3	pregnancy	0,236
cocoa	3	prenatal	0,227
pressure	3	labor	0,110
obstetrics	3	studies	0,106
risk	3	folate	0,090
cleft	3	premature	0,089

Table 5: Ranking of a vertex using a) betweenness and b) cluster. coeff.

Vertices	Betweenness Centrality	Vertices	Clustering Coefficient
vitamin	1117,5	prenatal	1
labor	834	malformation	1
risk	798	risk	0,333
pregnancy	487,5	premature	0,333
supplements	470,5	cleft	0,333
women	420	supplements	0,166
folate	323	orofacial	0,166
baby	270	labor	0,066
studies	171	vitamin	0,015
premature	58	–	–
incidence	58	–	–
malformation	24	–	–
cleft	20	–	–
orofacial	15	–	–
obstetrics	15	–	–
models	10	–	–
control	10	–	–
cocoa	9	–	–

Table 6: Linguistic blocks extracted from each retrieved sentence.

	Agent - NS	Patient - VS				
P	subject who?	verb	compl.1 what?	compl.2 for whom?	compl.3 where?	compl.4 when?
1	vitamin A	reduces	the risk of anemia	...	...	...
2	vitamin A	does not help prevent	risk of premature labor	...	...	...
3	levothyroxine	reduces	risk of premature labor	...	...	...
4	vitamin D + calcium	increases	risk of premature labor	...	...	...
5	liver oil	produces	vitamin D	...	...	...
6	mushroom	produces	vitamin D	...	...	...

passive to the active voice. For the compilation of Table 6, one must select all the verbs present in these sentences and then ask some questions (who? what? for whom? where? when?). These questions reveal the subjects/nouns and the verbal complements such as direct/indirect objects and accessory terms, i.e. verbal adjuncts of each sentence. For more information on the Verbka process, the reader is referred to [28].

With this set of retrieved sentences structured in tables, we move on to the construction of the concept map (Fig. 2). According to Novak [98] and Novak and Cañas [98], an important criterion for the construction of concept maps is the ideal number of key concepts needed. This number must vary from 15 to 25 concepts to facilitate the reading and understanding of the map by humans.

In this network, the subjects and the verbal complements of each sentence are called concepts, and they constitute the vertices. The relationship between two vertices is given by either a verb or by a preposition, and are represented by an edge (or directed arrow). This arrow starts from the subject concept and reaches the object concept(s) (Fig. 2). The key concepts are highlighted in the network in orange. The concepts highlighted in yellow are the ones considered to be most important from a first qualitative interpretation of the network, since they have a strong relationship with the key concepts.

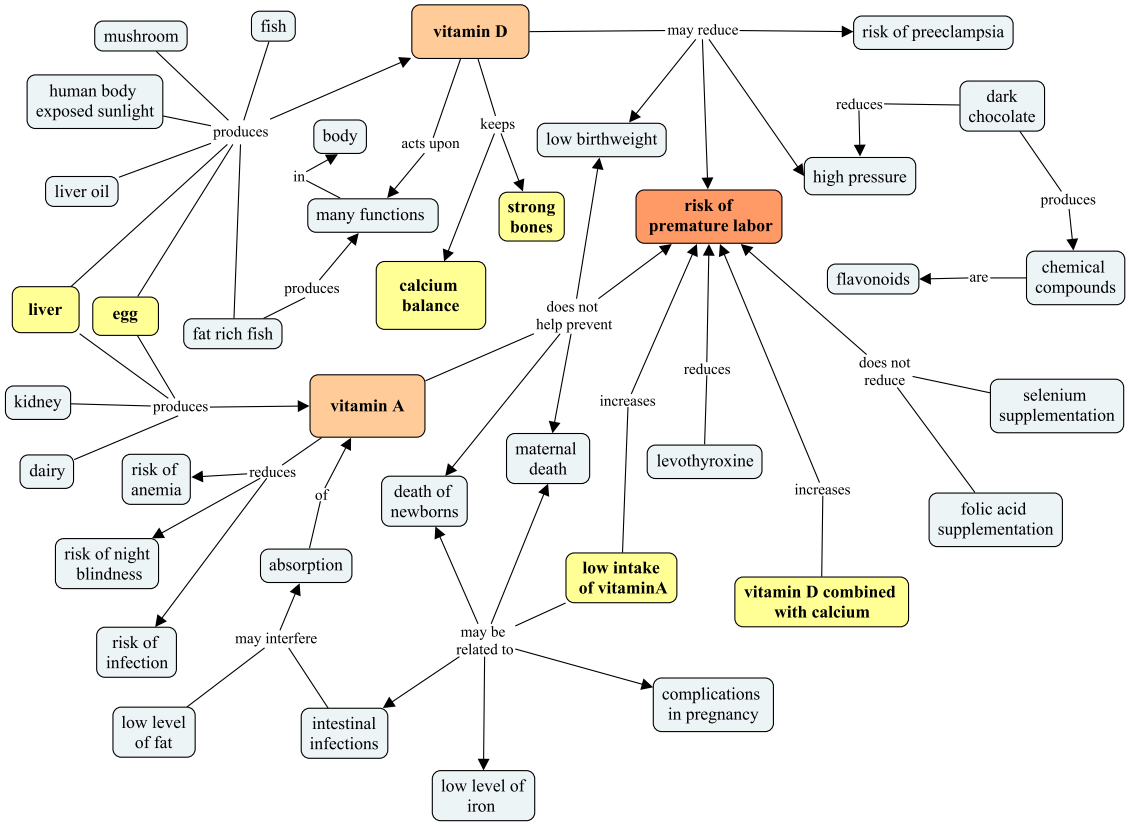

Figure 2: Concept map of propositions obtained from Table 6.

With this first constructed network model, we move on to the semantic classification of verbs to construct a causal concept map (Fig. 3). Thus, each of the edges was colored as follows: red for cause and effect relationships, blue for reflective relationships and black for static ones (i.e. those with no causal affectation). We emphasize that the network obtained in this application did not present reflexive verbal relationships (i.e. blue edges).

The information presented as cause and effect allowed the identification, through inferences made by the researcher, of (at least) four implicit relationships, as shown in Fig. 3 and 4 through dashed edges/arrows:

1. *Implicit relationship 1 (Fig. 3): "vitamin D combined with calcium may harm calcium balance"* .

   Given that a sufficient calcium balance has already been attained through vitamin D (which in turn is obtained from produces such as fish, mushroom, liver oil, and sunlight exposure), any additional supplement of vitamin D with calcium is likely to create an imbalance of calcium.

2. *Implicit relationship 2 (Fig. 3): "Vitamin D combined with calcium may generate excess calcium"*. This is a variant of the first implicit relationship, where the imbalance is a surplus. Since vitamin D keeps calcium, its combination with a calcium supplement may exacerbate the levels of calcium.

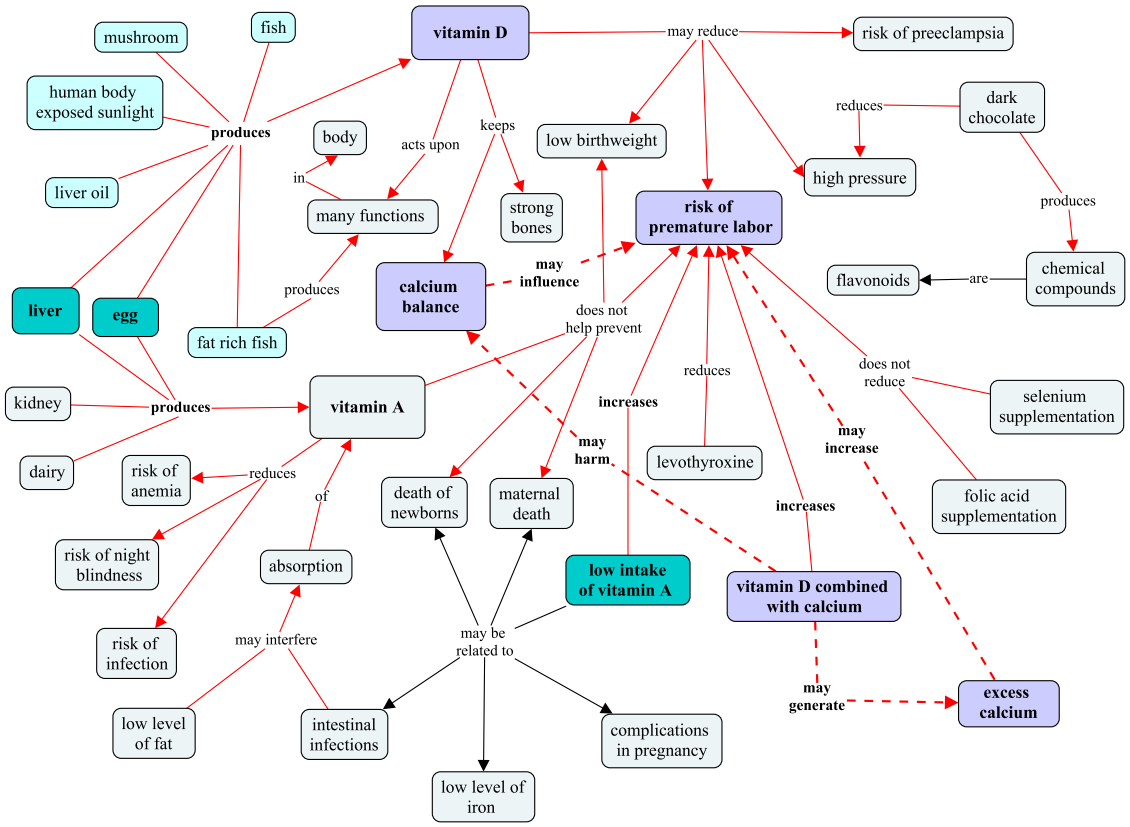

Figure 3: Causal concept map with inferences related to the effects caused by the ingestion of vitamin D supplements combined with calcium and its relationship with the increase in the risk of premature birth.

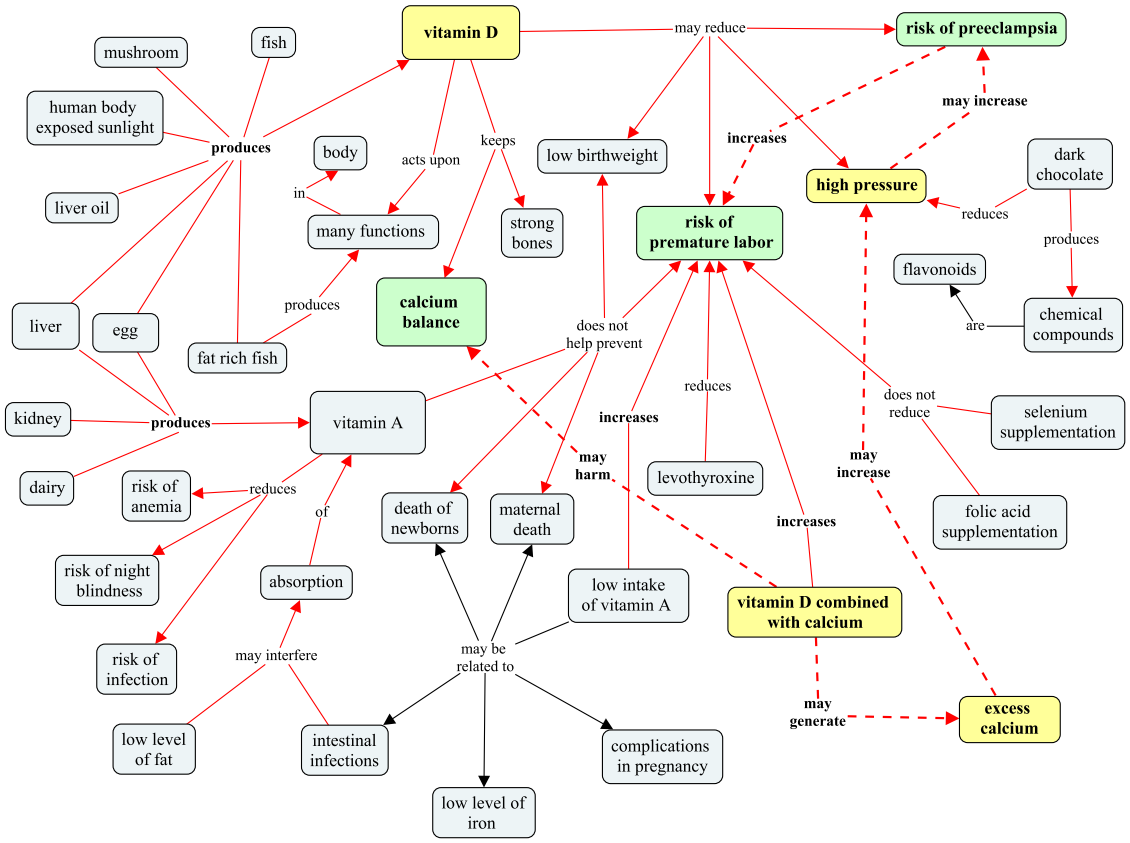

Figure 4: Causal concept map with inferences related to the impact of the excess of calcium on the blood (arterial) pressure.

3. *Implicit relationship 3 (Fig. 3): "Calcium balance may influence the risk of premature labor"*. Given that vitamin D may reduce the risk of premature labor whilst also keeping the calcium balance, one may formulate the hypothesis that vitamin D may reduce the risk of premature labor through the levels of calcium. Under this hypothesis, it may be that vitamin D is the direct and immediate cause, whereas vitamin D is the trigger.

4. *Implicit relationship 4 (Fig. 3):"Excess calcium may increase risk of premature labor"*. This is a variant of the preceding relationship, where the levels of calcium are in excess, indicating the type of influence poses a risk to premature labor.

5. *Implicit relationship 5 (Fig. 4): "Excess calcium may increase high (blood) pressure"*. The same situation previously mentioned regarding the imbalance of calcium in the organism, the excess of intake of this nutrient may also, in this case, affect the arterial pressure. Since vitamin D may reduce the risk of preeclampsia (a verified explicit relationship); and since vitamin D keeps the balance of calcium in the body (another verified explicit relationship); and since vitamin D may reduce blood pressure (the third explicit and verified relationship), therefore, the implicit relationship is that, excess of calcium may (as an hypothesis) increase blood pressure. Under this hypothesis, vitamin D affects the balance of calcium which then increases the blood pressure. Thus, vitamin D may be directly affecting the levels of calcium, and indirectly affecting the blood pressure through the excess of calcium.

6. *Implicit relationship 6 (Fig. 4): "High pressure may increase risk of preeclampsia"*. Whereas the map itself does not provide any solid clue about this relationship, we may establish it since it is known amongst practitioners in the field of medicine that high blood pressure during pregnancy is one of the biggest red flags that preeclampsia may be developing [2].

7. *Implicit relationship 7 (Fig. 4): "Risk of preeclampsia increases the risk of premature labor"*. Similarly, whereas the map itself provides no substantial evidence on this relationship, we may illustrate it in the causal concept map since it is known by domain experts that gestational hypertension-preeclampsia is the most common medical disorder of pregnancy [3].

As new relationships are found, these can be added to the table previously created (Table 6), thus increasing the amount of textual information and knowledge, now structured in rows and columns. Thus, the knowledge bases can constantly expand (Table 7, lines 7-13) each time a new implicit relationship is identified.

Table 7: Table of propositions with added implicit relationships (line 7 to 13) extracted from the causal concept map.

	Agent - NS	Patient - VS				
P	subject who?	verb	compl.1 what?	compl.2 for whom?	compl.3 where?	compl.4 when?
1	vitamin A	reduces	the risk of anemia	...	...	...
2	vitamin A	does not help prevent	risk of premature labor	...	...	...
3	levothyroxine	reduces	risk of premature labor	...	...	...
4	vitamin D + calcium	increases	risk of premature labor	...	...	...
5	liver oil	produces	vitamin D	...	...	...
6	mushroom	produces	vitamin D	...	...	...
7	vitamin D + calcium	may harm	calcium balance	...	...	...
8	vitamin D + calcium	may generate	excess calcium	...	...	...
9	calcium balance	may influence	risk of premature labor	...	...	...
10	excess calcium	may increase	risk of premature labor	...	...	...
11	excess calcium	may increase	high (blood) press	...	...	...
12	high pressure	may increase	risk of preeclampsia	...	...	...
13	risk of preeclampsia	may increase	risk of premature labor	...	...	...

Note that these implicit relationships are hypotheses that need to be corroborated by further scientific investigation. Nevertheless, the inference through the map was able to elicit these relationships. This example was chosen for its simplicity as a didactic way of illustrating the process. Clearly, other implicit relationships may be uncovered that may be more difficult to grasp using a simple and conventional approach (e.g. through reading a text). Notice too that, at best, the uncovering of this implicit relationship may be relatively easy obtained through the proposed approach (mainly once the tools are all integrated), in comparison to a conventional approach that would require the analyst to read more than 5000 words spread into ten different texts to achieve a similar result. At worst, either the analyst would

[2]https://www.preeclampsia.org/health-information/sign-symptoms
[3]https://www.ncbi.nlm.nih.gov/pubmed/16549208

not complete this task within a reasonable time, or would not accomplish it at all given the volume of information. Thus, we argue that the approach proposed in this work reached the proposed goal: to extract implicit information through text mining techniques and complex networks in conjunction with an approach based on verbal semantics.

5 Summary and Conclusion

Currently, one of the major challenges in the field of Knowledge Discovery in Text is the exploration of the large amount of information available in textual databases in an attempt to extract meaningful knowledge. Another challenge is finding the relationships between non-taxonomic concepts (based on events), which shows how these concepts affect each other.

In this work, we presented an Text Mining approach that attempts to address these concerns by being capable of generating a causal concept map from a textual corpus. The approach was based on the extraction and interpretation of concepts and their relationships. By mapping the explicit relationships onto a new view space such as a causal concept map, it was possible to reveal important characteristics of the set of texts that were used to extract implicit relationships. We chose to develop a process that facilitated the visualization and observation of the main explicit relationships existing in a given dataset and, consequently, aided in the inference of new relationships that were implicit in the same set. This was made possible due to the techniques used to decrease the number of words and the application of a semantic approach that created a new set of higher-level features from the original data.

The process handles the various stages of the Text Mining process, from data preprocessing to post-processing and the use of the results. With the application of association analysis in the data mining phase, it was possible to extract the most interesting rules from the set of texts and decrease the number of words. With the rules obtained, we were able to build a complex network and automatically extract its metrics. Thus, it was possible to assign weights to the items (words) and rank them. With this ranking, we obtained a subset of items, further reducing its number, and thus eliminating characteristics that were irrelevant for the task of extracting implicit relationships.

The process was satisfactory in dealing with two major challenges of text mining: the complexity and heterogeneity of textual data (word diversity) and the large number of its objects (words). With the reduction of the number of words (using association rules and complex networks), we were able to select a subset (key concepts) to generate a causal concept map. This map expresses the relationships obtained through the application of an analysis based on verbal semantics.

With the reduction of the number of words, we obtained a more understandable model, since this reduction allowed better visualization of the data in a concept map. This map contributed to the visualization of explicit relationships and enabled us to analyze a relatively large amount of contextual information, thus allowing the uncovering of relationships, by the user, that were implicit in the map.

To summarize, one problem that is circumvented in our approach is the potentially large number of patterns extracted by the algorithms in general, which works against the comprehensibility of user knowledge. The combined use of association rules and complex networks work both to mitigate this problem. It is possible to eliminate spurious patterns (false relationships) by applying association rules. Similarly, it is possible to curb the resulting large number of patterns by applying the metrics from complex networks. Another key task is the visualization of results. The concept map is a visualization model that can be used to facilitate the understanding and evaluation of the obtained result. We emphasize that this approach can be generalized to other domains, as the applied method is not bound to use predefined lists of terms or domain ontologies. The shortcoming of the method is the time taken to apply the semantic approach, which is performed manually.

Future work will focus on the automation of the knowledge acquisition process based on verbal semantics. This would allow the exploration of larger knowledge bases. Furthermore, other metrics may also be tried, including the analysis of nodes in the network with a lower degree. We also aim at performing a quantitative comparison against other methods.

6 Acknowledgments

This work was partially supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Grant Code 001.

References

[1] W. W. Fleuren and W. Alkema, “Application of text mining in the biomedical domain,” *Methods*, vol. 74, pp. 97–106, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.ymeth.2015.01.015

[2] H. Hashimi, A. Hafez, and H. Mathkour, “Selection criteria for text mining approaches,” *Computers in Human Behavior*, vol. 51, pp. 729–733, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.chb.2014.10.062

[3] R. Feldman and I. Dagan, "Knowledge discovery in textual databases (kdt)." in *KDD*, vol. 95, 1995, pp. 112–117.

[4] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, *Advances in knowledge discovery and data mining*. AAAI press Menlo Park, 1996, vol. 21.

[5] A. Usai, M. Pironti, M. Mital, and C. Aouina Mejri, "Knowledge discovery out of text data: a systematic review via text mining," *Journal of Knowledge Management*, vol. 22, no. 7, pp. 1471–1488, 2018. [Online]. Available: http://dx.doi.org/10.1108/JKM-11-2017-0517

[6] A. M. Cohen, C. E. Adams, J. M. Davis, C. Yu, P. S. Yu, W. Meng, L. Duggan, M. McDonagh, and N. R. Smalheiser, "Evidence-based medicine, the essential role of systematic reviews, and the need for automated text mining tools," in *Proceedings of the 1st ACM International Health Informatics Symposium*. ACM, 2010, pp. 376–380. [Online]. Available: http://dx.doi.org/10.1145/1882992.1883046

[7] J. Vashishtha, D. Kumar, and S. Ratnoo, "Revisiting interestingness measures for knowledge discovery in databases," in *Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on*. IEEE, 2012, pp. 72–78. [Online]. Available: http://dx.doi.org/10.1109/ACCT.2012.97

[8] A. Tan *et al.*, "Text mining: The state of the art and the challenges," in *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*, vol. 8. sn, 1999, pp. 65–70.

[9] H. Chen, *Knowledge management systems: a text mining perspective*. Knowledge Computing Corporation, 2001.

[10] S. O. Rezende, *Intelligent Systems: Fundamental Principles and Applications (in Portuguese)*. Editora Manole Ltda, 2003.

[11] S. A. Salloum, M. Al-Emran, A. A. Monem, and K. Shaalan, "Using text mining techniques for extracting information from research articles," in *Intelligent natural language processing: Trends and Applications*. Springer, 2018, pp. 373–397. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-67056-0_18

[12] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining, (First Edition)*. USA: Addison-Wesley Longman Publishing Co., Inc., 2005. [Online]. Available: http://dx.doi.org/10.5555/1095618

[13] B. J. Wielinga, "Reflections on 25+ years of knowledge acquisition," *International Journal of Human-Computer Studies*, vol. 71, no. 2, pp. 211–215, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.ijhcs.2012.10.007

[14] V. Sabol, W. Kienreich, M. Muhr, W. Klieber, and M. Granitzer, "Visual knowledge discovery in dynamic enterprise text repositories," in *Information Visualisation, 2009 13th International Conference*. IEEE, 2009, pp. 361–368. [Online]. Available: http://dx.doi.org/10.1109/IV.2009.35

[15] A. Singh, "A framework to automatically categorize the unstructured text documents," *Indian Journal of Science and Technology*, vol. 10, no. 8, 2017. [Online]. Available: http://dx.doi.org/10.17485/ijst/2017/v10i8/109472

[16] P. Bhardwaj and P. Khosla, "Review of text mining techniques," *IITM Journal of Management and IT*, vol. 8, no. 1, pp. 27–31, 2017.

[17] S.-J. Yang, N. N. Mishra, A. Rubio, and A. S. Bayer, "Causal role of single nucleotide polymorphisms within the mprf gene of staphylococcus aureus in daptomycin resistance," *Antimicrobial agents and chemotherapy*, pp. AAC–01 184, 2013. [Online]. Available: http://dx.doi.org/10.1128/AAC.01184-13

[18] S. Nirenburg and V. Raskin, *Ontological semantics*. Mit Press, 2004. [Online]. Available: http://dx.doi.org/10.1162/0891201053630246

[19] C. S. G. Khoo and J. C. Na, "Semantic relations in information science," *Annual Review of Information Science and Technology*, pp. 157–228, 2007. [Online]. Available: http://dx.doi.org/10.1002/aris.1440400112

[20] A. Kuhn, S. Ducasse, and T. Gírba, "Semantic clustering: Identifying topics in source code," *Information and Software Technology*, vol. 49, no. 3, pp. 230–243, 2007. [Online]. Available: http://dx.doi.org/10.1016/j.infsof.2006.10.017

[21] C. C. Aggarwal and C. X. Zhai, "A survey of text classification algorithms," in *Mining text data*. Springer, 2012, pp. 163–222. [Online]. Available: http://dx.doi.org/10.1007/978-1-4614-3223-4_6

[22] A. Huang, D. Milne, E. Frank, and I. H. Witten, "Clustering documents with active learning using wikipedia," in *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 2008, pp. 839–844. [Online]. Available: http://dx.doi.org/10.1109/ICDM.2008.80

[23] X. Wang, Y. Jia, R. Chen, H. Fan, and B. Zhou, "Improving text categorization with semantic knowledge in wikipedia," *IEICE TRANSACTIONS on Information and Systems*, vol. 96, no. 12, pp. 2786–2794, 2013. [Online]. Available: http://dx.doi.org/10.1587/transinf.E96.D.2786

[24] J. C. Trueswell and A. E. Kim, "How to prune a garden path by nipping it in the bud: Fast priming of verb argument structure," *Journal of memory and language*, vol. 39, no. 1, pp. 102–123, 1998. [Online]. Available: http://dx.doi.org/10.1006/jmla.1998.2565

[25] J. E. Boland, M. K. Tanenhaus, and S. M. Garnsey, "Evidence for the immediate use of verb control information in sentence processing," *Journal of Memory and Language*, vol. 29, no. 4, pp. 413–432, 1990. [Online]. Available: http://dx.doi.org/10.1016/0749-596X(90)90064-7

[26] D. Lin and P. Pantel, "Dirt@ sbt@ discovery of inference rules from text," in *Proceedings of the seventh ACM SIGKDD International conference on Knowledge discovery and data mining*. ACM, 2001, pp. 323–328. [Online]. Available: http://dx.doi.org/10.1145/502512.502559

[27] F. Hogenboom, F. Frasincar, U. Kaymak, F. De Jong, and E. Caron, "A survey of event extraction methods from text for decision support systems," *Decision Support Systems*, vol. 85, pp. 12–22, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.dss.2016.02.006

[28] D. G. Vasques, A. C. Zambon, G. B. Baioco, and P. S. Martins, "An Approach to Knowledge Acquisition Based on Verbal Semantics," in *Hawaii International Conference on System Sciences (HICSS)*. IEEE, jan 2016, pp. 4144–4153. [Online]. Available: http://dx.doi.org/10.1109/HICSS.2016.514

[29] J. Cong and H. Liu, "Approaching human language with complex networks," *Physics of life reviews*, vol. 11, no. 4, pp. 598–618, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.plrev.2014.04.004

[30] T. Gong, L. Shuai, and Y. Wu, "Extending network approach to language dynamics and human cognition. comment on "approaching human language with complex networks" by cong and liu," *Physics of life reviews*, vol. 11, pp. 639–640, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.plrev.2014.05.002

[31] G. A. Wachs-Lopes and P. S. Rodrigues, "Analyzing natural human language from the point of view of dynamic of a complex network," *Expert Systems with Applications*, vol. 45, pp. 8–22, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2015.09.020

[32] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Acm sigmod record*, vol. 22, no. 2. ACM, 1993, pp. 207–216. [Online]. Available: http://dx.doi.org/10.1145/170035.170072

[33] Z. Zheng, R. Kohavi, and L. Mason, "Real world performance of association rule algorithms," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001, pp. 401–406. [Online]. Available: http://dx.doi.org/10.1145/502512.502572

[34] B. Padmanabhan and A. Tuzhilin, "Unexpectedness as a measure of interestingness in knowledge discovery," *Decision Support Systems*, vol. 27, no. 3, pp. 303–318, 1999. [Online]. Available: http://dx.doi.org/10.1016/S0167-9236(99)00053-6

[35] M. D. Ruiz, J. Gómez-Romero, M. Molina-Solana, M. Ros, and M. J. Martín-Bautista, "Information fusion from multiple databases using meta-association rules," *International Journal of Approximate Reasoning*, vol. 80, pp. 185–198, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.ijar.2016.09.006

[36] C. d'Amato, S. Staab, A. G. Tettamanzi, T. D. Minh, and F. Gandon, "Ontology enrichment by discovering multi-relational association rules from ontological knowledge bases," in *Proceedings of the 31st Annual ACM Symposium on Applied Computing*. ACM, 2016, pp. 333–338. [Online]. Available: http://dx.doi.org/10.1145/2851613.2851842

[37] A. Amin, R. Talib, S. Raza, and S. Javed, "Extract association rules to minimize the effects of dengue by using a text mining technique," *International Journal of Computer Science and Mobile Computing*, vol. 3, no. 4, pp. 394–400, 2014.

[38] D. G. Vasques, L. Comelli, C. M. Ranieri, and S. O. Rezende, "Text mining for client management in telecom companies (in portuguese)," in *XVIII Brazilian symposium of information systems*, 2017, pp. 9–16.

[39] H. Jiang, H. T. Horner, T. M. Pepper, M. Blanco, M. Campbell, and J.-l. Jane, "Formation of elongated starch granules in high-amylose maize," *Carbohydrate Polymers*, vol. 80, no. 2, pp. 533–538, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.carbpol.2009.12.016

[40] R. G. Rossi, "Automatic classification of texts by means of network-based machine learning (in portuguese)," Ph.D. dissertation, Universidade de São Paulo, 2016. [Online]. Available: http://dx.doi.org/10.11606/T.55.2016.tde-05042016-105648

[41] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, p. 47, 2002. [Online]. Available: http://dx.doi.org/10.1103/RevModPhys.74.47

[42] X. Ke, Y. Zeng, Q. Ma, and L. Zhu, "Complex dynamics of text analysis," *Physica A: Statistical Mechanics and its Applications*, vol. 415, pp. 307–314, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.physa.2014.08.022

[43] Q. Lu and L. Getoor, "Link-based classification," in *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 2003, pp. 496–503.

[44] R. V. Solé, B. Corominas-Murtra, S. Valverde, and L. Steels, "Language networks: Their structure, function, and evolution," *Complexity*, vol. 15, no. 6, pp. 20–26, 2010. [Online]. Available: http://dx.doi.org/10.1002/cplx.20305

[45] Z. Xu, X. Wei, X. Luo, Y. Liu, L. Mei, C. Hu, and L. Chen, "Knowle: a semantic link network based system for organizing large scale online news events," *Future Generation Computer Systems*, vol. 43, pp. 40–50, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.future.2014.04.002

[46] J. F. Sowa, *Conceptual Structures: Information Processing in Mind and Machine*. USA: Addison-Wesley Longman Publishing Co., Inc., 1984.

[47] D. Allemang and J. Hendler, *Semantic web for the working ontologist: effective modeling in RDFS and OWL*. Elsevier, 2011. [Online]. Available: http://dx.doi.org/10.1016/C2010-0-68657-3

[48] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, and T. C. Rindflesch, "Semmeddb: a pubmed-scale repository of biomedical semantic predications," *Bioinformatics*, vol. 28, no. 23, pp. 3158–3160, 2012. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/bts591

[49] T. C. Rindflesch and M. Fiszman, "The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text," *Journal of biomedical informatics*, vol. 36, no. 6, pp. 462–477, 2003. [Online]. Available: http://dx.doi.org/10.1016/j.jbi.2003.11.003

[50] H. Kilicoglu, M. Fiszman, A. Rodriguez, D. Shin, A. Ripple, and T. C. Rindflesch, "Semantic medline: a web application for managing the results of pubmed searches," in *Proceedings of the third International symposium for semantic mining in biomedicine*, vol. 2008. Citeseer, 2008, pp. 69–76.

[51] T. C. Rindflesch, H. Kilicoglu, M. Fiszman, G. Rosemblat, and D. Shin, "Semantic medline: An advanced information management application for biomedicine," *Information Services & Use*, vol. 31, no. 1-2, pp. 15–21, 2011. [Online]. Available: http://dx.doi.org/10.3233/ISU-2011-0627

[52] G. Hodge, *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files.* ERIC, 2000.

[53] L. Café and M. Bräscher, "Knowledge organization: semantic theories as a basis for the study and representation of concepts (in portuguese)," *Informação & Informação*, vol. 16, no. 2, 2011. [Online]. Available: http://dx.doi.org/10.5433/1981-8920.2011v16n2p25

[54] H. Gardner, *The Mind's New Science: A History of the Cognitive Revolution*. USA: Basic Books, Inc., 1985.

[55] P. Pantel and D. Lin, "A statistical corpus-based term extractor," in *Conference of the Canadian Society for Computational Studies of Intelligence*. Springer, 2001, pp. 36–46. [Online]. Available: http://dx.doi.org/10.1007/3-540-45153-6_4

[56] T. M. Chung, "A corpus comparison approach for terminology extraction," *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, vol. 9, no. 2, pp. 221–246, 2003. [Online]. Available: http://dx.doi.org/doi.org/10.1075/term.9.2.05chu

[57] N. Guarino and C. Welty, "Evaluating ontological decisions with ontoclean," *Communications of the ACM*, vol. 45, no. 2, pp. 61–65, 2002. [Online]. Available: http://dx.doi.org/10.1145/503124.503150

[58] A. Schutz and P. Buitelaar, "Relext: A tool for relation extraction from text in ontology extension," in *International semantic web conference*. Springer, 2005, pp. 593–606. [Online]. Available: http://dx.doi.org/10.1007/11574620_43

[59] D. SáncLearning nonhez and A. Moreno, "Learning non-taxonomic relationships from web documents for domain ontology construction," *Data & Knowledge Engineering*, vol. 64, no. 3, pp. 600–623, 2008. [Online]. Available: http://dx.doi.org/10.1016/j.datak.2007.10.001

[60] M. Kavalec, A. Maedche, and V. Svátek, "Discovery of lexical entries for non-taxonomic relations in ontology learning," in *International Conference on Current Trends in Theory and Practice of Computer Science*. Springer, 2004, pp. 249–256. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-24618-3_21

[61] C. Fillmore, "The case for case," *In E. Bach & RT Harms (eds.). Universals in Linguistic Theory. New York: Holt Rinehart and Winston*, pp. 1–88, 1968.

[62] S. C. Dik, *Functional grammar*. Foris Pubns USA, 1981, vol. 7.

[63] T. Givón, "Direct object and dative shifting: Semantic and pragmatic case," *Objects: Towards a theory of grammatical relations*, pp. 151–182, 1984.

[64] J. Bresnan and J. M. Kanerva, "Locative inversion in chichewa: a case study of factorization in grammar," *Linguistic inquiry*, pp. 1–50, 1989. [Online]. Available: http://dx.doi.org/10.1163/9789004373181_006

[65] R. Jackendoff, "On larson's treatment of the double object construction," *Linguistic inquiry*, vol. 21, no. 3, pp. 427–456, 1990.

[66] J. R. D. VanValin, "Semantic parameters of split intransitivity," *Language*, pp. 221–260, 1990.

[67] H. Llorens, E. Saquete, and B. Navarro-Colorado, "Applying semantic knowledge to the automatic processing of temporal expressions and events in natural language," *Information Processing & Management*, vol. 49, no. 1, pp. 179–197, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.ipm.2012.05.005

[68] D. Garcia *et al.*, "Coatis, an nlp system to locate expressions of actions connected by causality links," in *International Conference on Knowledge Engineering and Knowledge Management*. Springer, 1997, pp. 347–352. [Online]. Available: http://dx.doi.org/10.1007/BFb0026799

[69] C. S. Khoo, J. Kornfilt, R. N. Oddy, and S. H. Myaeng, "Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing," *Literary and Linguistic Computing*, vol. 13, no. 4, pp. 177–186, 1998. [Online]. Available: http://dx.doi.org/10.1093/llc/13.4.177

[70] R. Girju, "Automatic detection of causal relations for question answering," in *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*. Association for Computational Linguistics, 2003, pp. 76–83. [Online]. Available: http://dx.doi.org/10.3115/1119312.1119322

[71] V. Nastase, J. Sayyad-Shirabad, M. Sokolova, and S. Szpakowicz, "Learning noun-modifier semantic relations with corpus-based and wordnet-based features," in *AAAI*, 2006, pp. 781–787. [Online]. Available: http://dx.doi.org/10.5555/1597538.1597663

[72] A. S. H. Al Hashimy and N. Kulathuramaiyer, "Ontology enrichment with causation relations," in *Systems, Process & Control (ICSPC), 2013 IEEE Conference on*. IEEE, 2013, pp. 186–192. [Online]. Available: http://dx.doi.org/10.1109/SPC.2013.6735129

[73] J. Todhunter, I. Sovpel, and H. Zhyhalko, "System and method for cross-language knowledge searching," 2010, uS Patent 7,672,831.

[74] D. G. Vasques, F. D. Gomes, J. F. G. Jaramillo, and P. S. Martins, "Verbka: An Approach To Building Causal Concept Maps Based On Verbal Semantics," in *7th International Conference on Concept Mapping*, Tallinn, 2016.

[75] R. W. Langacker, *Foundations of cognitive grammar: Theoretical prerequisites*. Stanford university press, 1987, vol. 1.

[76] D. Dowty, "Thematic proto-roles and argument selection," *language*, vol. 67, no. 3, pp. 547–619, 1991. [Online]. Available: http://dx.doi.org/10.1353/lan.1991.0021

[77] X. Liu, X. Zhang, Y. Wang, J. Zhou, S. Helal, Z. Xu, W. Zhang, and S. Cao, "Parmtrd: Parallel association rules based multiple-topic relationships detection," in *International Conference on Web Services*. Springer, 2018, pp. 422–436. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-94289-6_27

[78] K. Zupanc and J. Davis, "Estimating rule quality for knowledge base completion with the relationship between coverage assumption," in *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 2018, pp. 1073–1081. [Online]. Available: http://dx.doi.org/10.1145/3178876.3186006

[79] J. Kralj, M. Robnik-Sikonja, and N. Lavrac, "Netsdm: Semantic data mining with network analysis." *Journal of Machine Learning Research*, vol. 20, no. 32, pp. 1–50, 2019.

[80] A. E. Sizemore, E. A. Karuza, C. Giusti, and D. S. Bassett, "Knowledge gaps in the early growth of semantic feature networks," *Nature human behaviour*, vol. 2, no. 9, p. 682, 2018. [Online]. Available: http://dx.doi.org/10.1038/s41562-018-0422-4

[81] J. Raimbault, "Exploration of an interdisciplinary scientific landscape," *Scientometrics*, vol. 119, no. 2, pp. 617–641, 2019. [Online]. Available: http://dx.doi.org/10.1007/s11192-019-03090-3

[82] C. Yang, C. Huang, and J. Su, "An improved sao network-based method for technology trend analysis: A case study of graphene," *Journal of Informetrics*, vol. 12, no. 1, pp. 271–286, 2018. [Online]. Available: http://dx.doi.org/10.1016/j.joi.2018.01.006

[83] W. J. Vlietstra, R. Zielman, R. M. van Dongen, E. Schultes, F. Wiesman, R. Vos, E. M. van Mulligen, and J. A. Kors, "Automated extraction of potential migraine biomarkers using a semantic graph," *Journal of biomedical informatics*, vol. 71, pp. 178–189, 2017.

[84] J. Qiu, Y. Chai, Y. Liu, Z. Gu, S. Li, and Z. Tian, "Automatic non-taxonomic relation extraction from big data in smart city," *IEEE Access*, vol. 6, pp. 74 854–74 864, 2018. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2018.2881422

[85] K. Yu, P.-Y. Lung, T. Zhao, P. Zhao, Y.-Y. Tseng, and J. Zhang, "Automatic extraction of protein-protein interactions using grammatical relationship graph," *BMC medical informatics and decision making*, vol. 18, no. 2, p. 42, 2018. [Online]. Available: http://dx.doi.org/10.1186/s12911-018-0628-4

[86] A. Lugowski, S. Kamil, A. Buluç, S. Williams, E. Duriakova, L. Oliker, A. Fox, and J. R. Gilbert, "Parallel processing of filtered queries in attributed semantic graphs," *Journal of Parallel and Distributed Computing*, vol. 79, pp. 115–131, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.jpdc.2014.08.010

[87] Y. Bocanegra, A. M. García, D. Pineda, O. Buriticá, A. Villegas, F. Lopera, D. Gómez, C. Gómez-Arias, J. F. Cardona, N. Trujillo *et al.*, "Syntax, action verbs, action semantics, and object semantics in parkinson's disease: Dissociability, progression, and executive influences," *Cortex*, vol. 69, pp. 237–254, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.cortex.2015.05.022

[88] D. C. Corrales, A. Ledezma, and J. C. Corrales, "A conceptual framework for data quality in knowledge discovery tasks (fdq-kdt): A proposal." *JCP*, vol. 10, no. 6, pp. 396–405, 2015. [Online]. Available: http://dx.doi.org/10.17706/jcp.10.6.396-405

[89] N. F. Ebecken, M. C. S. Lopes, M. C. Costa *et al.*, "Text mining (in portuguese)," *Sistemas inteligentes: fundamentos e aplicações. São Carlos: Manole*, pp. 337–370, 2003.

[90] J. L. G. Rosa, "The meaning of the word for natural language processing (in portuguese)," *Work presented at the ZLV Seminar of GEL, University of Campinas (Unicamp), Campinas*, 1997.

[91] ——, "A symbolic-connectionist hybrid system for processing thematic roles (in portuguese)," Ph.D. dissertation, University of Campinas (Unicamp). Institute of Language Studies, 1999.

[92] R. G. Rossi and S. O. Rezende, "Generating features from textual documents through association rules," *Proceedings of the National Meeting on Artificial Intelligence. SBC. Sao Carlos–SP–Brazil*, p. 20, 2011.

[93] A. A. Lopes, R. Pinho, F. V. Paulovich, and R. Minghim, "Visual text mining using association rules," *Computers & Graphics*, vol. 31, no. 3, pp. 316–326, 2007. [Online]. Available: http://dx.doi.org/10.1016/j.cag.2007.01.023

[94] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proceedings of the 20th International Conference on Very Large Data Bases*, ser. VLDB '94. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, p. 487–499. [Online]. Available: http://dx.doi.org/10.5555/645920.672836

[95] C. Zhang and S. Zhang, *Association rule mining: models and algorithms*. Springer-Verlag, 2002. [Online]. Available: http://dx.doi.org/10.5555/1791549

[96] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003. [Online]. Available: http://dx.doi.org/10.1137/S003614450342480

[97] M. Bastian, S. Heymann, M. Jacomy *et al.*, "Gephi: an open source software for exploring and manipulating networks." *Icwsm*, vol. 8, pp. 361–362, 2009.

[98] A. J. Cañas and J. D. Novak, "The theory underlying concept maps and how to construct and use them," *Práxis Educativa*, vol. 5, no. 1, pp. 9–29, 2010.

[99] J. D. Novak, *Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and corporations*. Routledge, 2010.